Next-Generation Global Biomonitoring

CellPress

Opinion

Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks

David A. Bohan,[1,*] Corinne Vacher,[2] Alireza Tamaddoni-Nezhad,[3] Alan Raybould,[4] Alex J. Dumbrell,[5] and Guy Woodward[6]

We foresee a new global-scale, ecological approach to biomonitoring emerging within the next decade that can detect ecosystem change accurately, cheaply, and generically. Next-generation sequencing of DNA sampled from the Earth's environments would provide data for the relative abundance of operational taxonomic units or ecological functions. Machine-learning methods would then be used to reconstruct the ecological networks of interactions implicit in the raw NGS data. Ultimately, we envision the development of autonomous samplers that would sample nucleic acids and upload NGS sequence data to the cloud for network reconstruction. Large numbers of these samplers, in a global array, would allow sensitive automated biomonitoring of the Earth's major ecosystems at high spatial and temporal resolution, revolutionising our understanding of ecosystem change.

Bioindicators for Change in Ecosystem Functioning

Environmental change is increasingly reshaping biodiversity and the provision of ecosystem processes and services across local to global scales [1,2]. Yet, we are poorly equipped to measure these relationships and rely on judgements drawn from proxies or **biomonitoring indicators** (see Glossary). Chemical indicators can evaluate some environmental stressors [3], but often these are transient and hard to measure directly and so biotic indicators are used to gauge impacts and responses [4,5]. Biomonitoring underpins many areas of policy [6] and, in the case of 'charismatic' indicator species, have considerable value for the public. In almost all cases, however, biomonitoring suffers from at least one of three key problems: (i) limited accuracy, because indicators are simple proxies that cannot capture the full range of complex ecological phenomena; (ii) high costs that limit the scale of coverage, especially in the majority of systems that rely on human labour for sampling rather than automation; and (iii) limited generality, because most are bespoke designs focused on specific systems and individual stressors. Most biomonitoring schemes use methods developed in the middle of the 20th century and not the approaches that have since appeared. Consequently, biomonitoring of the full diversity of species and their interactions within an ecosystem is rarely, if ever, attempted. We foresee that a new generation of biomonitoring will be needed in the next decade, to complement indicator-based approaches, which detects systemic change in any ecosystem more accurately, cheaply, and generically at local to global scales.

Trends

Next-generation sequencing (NGS) can used to sample nucleic acids in the environment for the presence of species and ecological functions.

Machine-learning software can search for 'the ghosts of interactions past' in the raw NGS data to reconstruct the networks of ecological interactions.

NGS data and machine-learning in the cloud could be combined in the next generation of global biomonitoring. Autonomous NGS samplers would sequence and upload data for ecological network reconstruction, to detect ecosystem change accurately, cheaply and generically.

Reconstruction of highly replicated networks of ecological interaction, using this next generation of biomonitoring, would provide general ecological information for ecosystem comparison and a revolution in the breadth of our understanding of the ecology of ecosystem change.

[1]Agroécologie, AgroSup Dijon, INRA, University of Bourgogne Franche-Comté, F-21000 Dijon, France
[2]BIOGECO, INRA, University of Bordeaux, 33615 Pessac, France
[3]Computational Bioinformatics Laboratory, Department of Computing, Imperial College London, London, SW7 2AZ, UK
[4]Syngenta Crop Protection AG, PO

Box 4002, Basel, Switzerland
[5]School of Biological Sciences, University of Essex, Colchester, Essex, CO4 3SQ, UK
[6]Department of Life Sciences, Imperial College London, Silwood Park Campus, Berkshire, SL5 7PY, UK

*Correspondence: David.Bohan@inra.fr (D.A. Bohan).

In our vision, **next-generation sequencing (NGS)** approaches are used to sample nucleic acids in the environment to quantify the abundance of species or **operational taxonomic units (OTUs)**, and/or the level of expression of key functional genes (Figure 1). Then, **cloud**-based, machine-learning will automatically reconstruct the ecological networks implicit in the sample data [7]. We envision the development of miniaturised, autonomous NGS samplers that can be deployed in large numbers across global arrays to provide standardised and sensitive automated sampling and remote biomonitoring of all the Earth's ecosystems at high resolution and in real time. By reconstructing highly replicated networks of ecological interactions, this biomonitoring would provide the global standard of ecosystem information and revolutionise our ability to measure, understand, and predict how the planet's ecosystems respond to environmental change.

Ecological Network Structure Determines Ecosystem Functioning

Ecological networks are increasingly being used to characterise the system-level responses to environmental change, including pollution, land use, and climate change, overexploitation, species invasions and disease [8–10]. All these stressors fall within the remit of biomonitoring. However, biomonitoring typically focuses on specific indicator **nodes** (species or OTUs) or **links** (functions or interactions) rather than the totality of nodes and links that bind the ecosystem together. Specific nodes and links can play keystone roles in networks and act as indicators of environmental stress. Network studies are increasingly demonstrating that ecosystem function and ecosystem responses to change are related to network structure (accuracy), and in ways that cannot be predicted from studying individual nodes and links in isolation [11–14].

The complex relationships between changes in nodes and links and their impact on ecosystem functions should be understood at the network level if we are to develop more robust biomonitoring. Environmental change triggers effects that propagate through an ecological

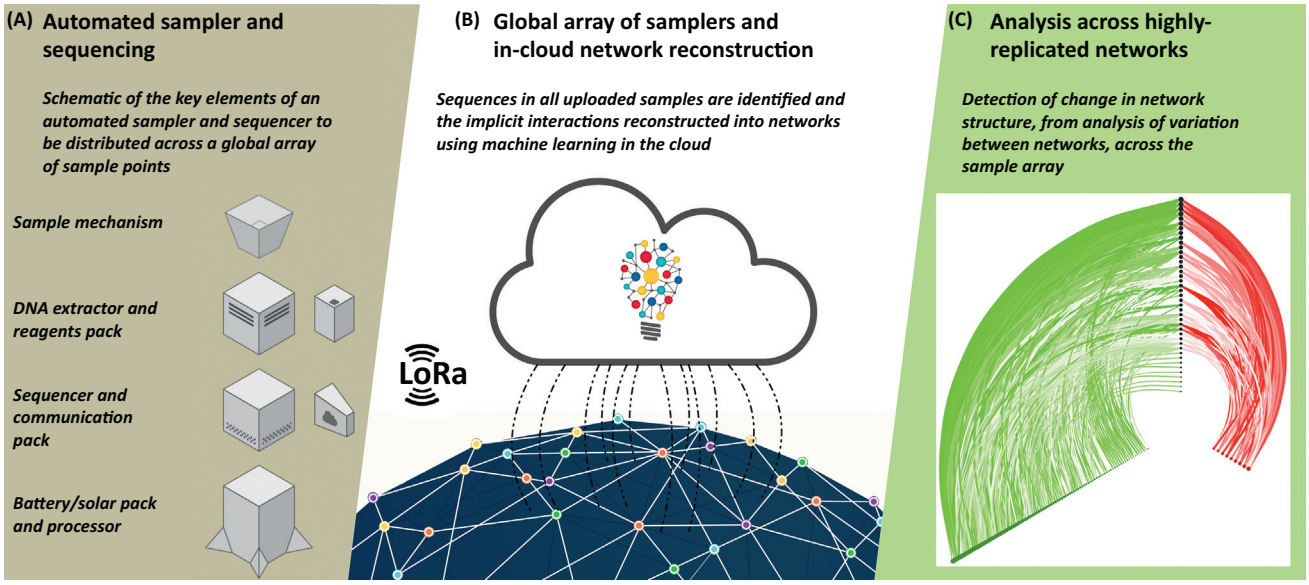

Figure 1. Large-Scale Biomonitoring Using NGS. Schematic of the workflow, from (A) to (C) for the NGB approach that we advocate. (A) Schematic illustration of the key components of an autonomous sampler. (B) Diagram of an array of sample points, each with a sampler, and the upload of sequence data to the cloud via the latest communication standards (here from the LoRa Alliance[ii]). Management, identification and reconstruction of network structure is done in the cloud. (C) Detection and analysis of change in the structure of the monitored networks.

network, modulating the ecosystem's response [15,16]. Without an understanding of top-down effects in food webs, for example, counterintuitive responses such as declines in invertebrate numbers under reduced environmental stress cannot be interpreted [17]. Network structure also explains rapid shifts between stable ecosystem states, even under otherwise identical environmental conditions [18,19], and why these modified states can persist long-term [20]. Ecological networks allow us to deal with some of the problems of generality that can bedevil biomonitoring. The theory argues for the structure of the network, rather than specific nodes or links, being key to ecosystem function [21]. Networks would therefore provide the generality of understanding of ecosystem function across systems and biogeographical regions needed for global biomonitoring tools.

While network approaches can help to improve on the problems of accuracy and generality, they cannot yet be used routinely in biomonitoring because they fail to satisfy problems of cost. Constructing ecological networks requires considerable time and manpower to observe and parameterise the linkages between the species in a system, even for small networks [22,23]. This cost has rendered the construction of networks too slow and the resource demands too high for their practicable use in biomonitoring.

New Tools to Accelerate Network Characterisation

Machine learning offers huge potential for reconstructing ecological networks from available data. Statistical inference and logic-based machine learning approaches have been widely used in the social sciences and for molecular and genetic interaction networks, but have only recently been applied in ecology (Box 1). The idea behind these machine-learning methods is simple; embedded in a dataset is the imprint of the recent processes and interactions that created the data, and this information can be recovered to reconstruct networks.

In an ecological dataset, variation is habitually used by ecologists to test hypotheses about past reproduction, migration and predation, for example using statistical regression. For machine learning, the use of the variation in the data is extended to state that these 'ghosts of interactions past' can be recovered to reconstruct ecological networks of interaction [7]. Machine-learning methods are a search for correlation or relational patterns in the data and thus have similarities to, and the same weaknesses of, many model-fitting approaches.

Box 1. Machine Learning Ecological Network Structure

The aim of reconstructing ecological networks is to recover the biotic interactions (e.g., competition, parasitism, mutualism) that structure communities from the observed variation in sample data. This can be done from ecological time series that record the temporal variation of ecological communities or from the spatial variation in species occurrence or abundance, including time-resolved [45–47] and spatially-resolved [48–50] metagenetic datasets. The learning uses two forms of approach to recover biotic interactions – statistical inference and logical machine learning – that constitute two independent and complementary lines of research. It is not yet clear which of these approaches will be most suited to biomonitoring ecological networks [7,51]. What is clear is that cloud-based learning, based on either of these methods, is possible and already in use. Learning approaches have for instance been used to develop language translations tools such as Google Translate.

The underlying hypothesis of network reconstruction is that biotic interactions produce correlations and relational patterns in the abundance of species. In the case of statistical inference, the variation in the sample is treated statistically. Significant correlations between the abundances of any two species within the dataset are considered as potential network edges. The challenge is to remove edges that are not likely to be caused by biotic interactions, by integrating background information such as environmental factors and species functional traits as covariates into the statistical model [7,36,52,53]. Logical machine learning considers relational patterns rather like the structure of grammar in a language [54–56]. The grammar of a trophic interaction, for example, can be coded as background information – in the agro-ecological network the predator and prey species must co-occur in the same samples, the predator should have appropriate mouthparts and predators should be larger than their prey. Trophic interactions between two species are only identified if this grammar rule is realized. For both statistical inference and logical approaches, such background knowledge can come from our current ecological knowledge or be learnt from observed data.

Glossary

Amplicon: a piece of DNA or RNA that is the product of natural or artificial (i.e., PCR) replication.
Biomonitoring: biological monitoring, or biomonitoring, uses biological responses to evaluate change, caused by pollution, climate, human management and conservation, species invasions and disease.
Cloud: a type of computing infrastructure that provides shared computer processing resources and data to computers and other devices on demand across the internet. The shared computing resources, such as computer networks, servers, storage, applications, and services, can be tailored to particular needs with minimal management effort. Users have the capability to store and process their data in third-party data centres, located anywhere in the world. The availability of high-capacity networks, low-cost computers, and storage devices have driven the growth in cloud computing.
Indicators: indicators are used in biomonitoring to evaluate risks to human health and the environment for communication to the public or government policy makers. Pesticides, elements and metabolites are commonly used as pollution indicators, while Species, communities and behavioural approaches are also used to infer the ecological condition of terrestrial and aquatic ecosystems. The development and using of indicators is essentially pragmatic. To date, the evaluation of the myriad changes in ecosystems that can occur is too costly and time intensive.
Link: a link, or edge, connects two nodes in a network. Information transacted across a link can be undirected (the flow goes both ways) or directed (one way). In the case of energy pathways, directed links represent energy flux. In the case of mutualistic networks, a pair of directed links represents an interaction with mutual benefit, such as in the case of plant-pollination. For classical food webs, directed links go from the prey/resource to the predator/consumer.
Next-generation sequencing (NGS): NGS, or high-throughput sequencing, allows the sequencing of DNA and RNA much more rapidly

Consequently, it is necessary to take care in choosing the learning method with a particular dataset, and to take into account the biases in all sample data; including taxa biases, identification errors, zero-rich data, and non-normal abundance distributions. When selected correctly, however, we find that machine learning methods have great power to rapidly reconstruct networks from ecological sample data, in principle permitting their use in biomonitoring.

Reconstructing a Large Agro-Ecological Network

To reconstruct a replicated, invertebrate food web from a large herbicide treatment dataset [24,25], Bohan *et al.* [26] first 'exposed' trophic interactions by transforming their data of species abundances to logarithmic treatment-ratios across two levels of an herbicide treatment. Their thinking was that following an application of herbicide, weed plants that provide refuge/food resources to a prey, *y*, would die. *y* would then either move or die, affecting the ratio. A predator of *y*, *x*, would in turn move or die and it might therefore be hypothesised that species undergoing trophic interactions would have correlated changes in their treatment ratios.

Guiding the Learning – Background Knowledge

Correlations in data arise for many reasons, including chance, and do not imply a trophic interaction. To maximise the likelihood of learning a predation interaction, the learning was guided using 'background information' that serves as a model for a trophic interaction. It posited that a trophic interaction is one in which: (i) predator *x* co-occurs at the same sample points as prey *y*; (ii) *x* has appropriate mouthparts to consume *y*; and (iii) calling on a basic hypothesis of trophic ecology, *x* should have a larger body size than *y* – big things eat small things [14,26].

Validating the Reconstructed Network

The agricultural network that was reconstructed, using a logic-based machine learning approach [27], bore all the hallmarks of an ecological food web [7]. Validation was done through an analysis of the literature and significant Pearson correlation was noted [26,28]. Moreover, the frequency of learnt links correlated well with the frequency the link was found in the literature [28,29]. In essence, the machine learning was reconstructing the network that might have been hypothesised from expert knowledge and the literature (see Figure 2).

New Understanding

The aim of machine learning is not to simply reproduce what we already know, but also to learn new science. In the agricultural network, apparently illogical links implicating spiders as prey were posited. This was unexpected and three possible explanations of the finding were proposed: either the learning was incorrect; or the small-bodied spiders were learnt as prey as an artefact of the hypothesis that 'big things eat small things'; or spiders do indeed serve as prey within this food web [26,28]. The third explanation was tested by examining the gut contents of stored samples of 'spider-predator' species using spider-specific DNA **primers**. Several species of spiders were subsequently shown to be present in the predator guts [30], demonstrating, via an inferred hypothesis and explicit test, the potential of machine learning to discover new science.

Network Reconstruction from Ubiquitous Data

This food web example shows that it is possible to quickly learn ecological networks. Therefore, once appropriately selected, machine learning methods can produce valid networks, do science and greatly reduce the costs of building networks over traditional means [26,28,29]. However, to use network learning for biomonitoring in any ecosystem, it is necessary to move away from classical, ecological sampling to a generic source of ecological data.

and cheaply than the prior technology of Sanger sequencing, thus 'revolutionising' genomics and molecular biology. It is a catch-all term to describe a number of different sequencing methodologies including Solexa (Illumina), Roche 454, Proton/PGM, PacBio, GridION/MinION and SOLiD sequencing.

Node: a node, or vertex, represents an individual component of a network, for example a species in a species–species interaction network like a food web or a plant–pollinator network. In an NGS reconstructed network, the OTUs are treated as nodes.

Operational taxonomic unit (OTU): a pragmatic definition for a group of closely related individuals. In this article we use the term OTU to refer to clusters of organisms, grouped by their sequence similarity for a specific set of taxonomic marker genes, such as the 16S or 18S rRNA genes.

Polymerase chain reaction (PCR): a method used in molecular biology to copy or replicate a single or a few pieces of DNA over several orders of magnitude of times, producing thousands to millions of copies of a particular DNA sequence.

Primer: a short strand of RNA or DNA that serves as a starting template for DNA replication.

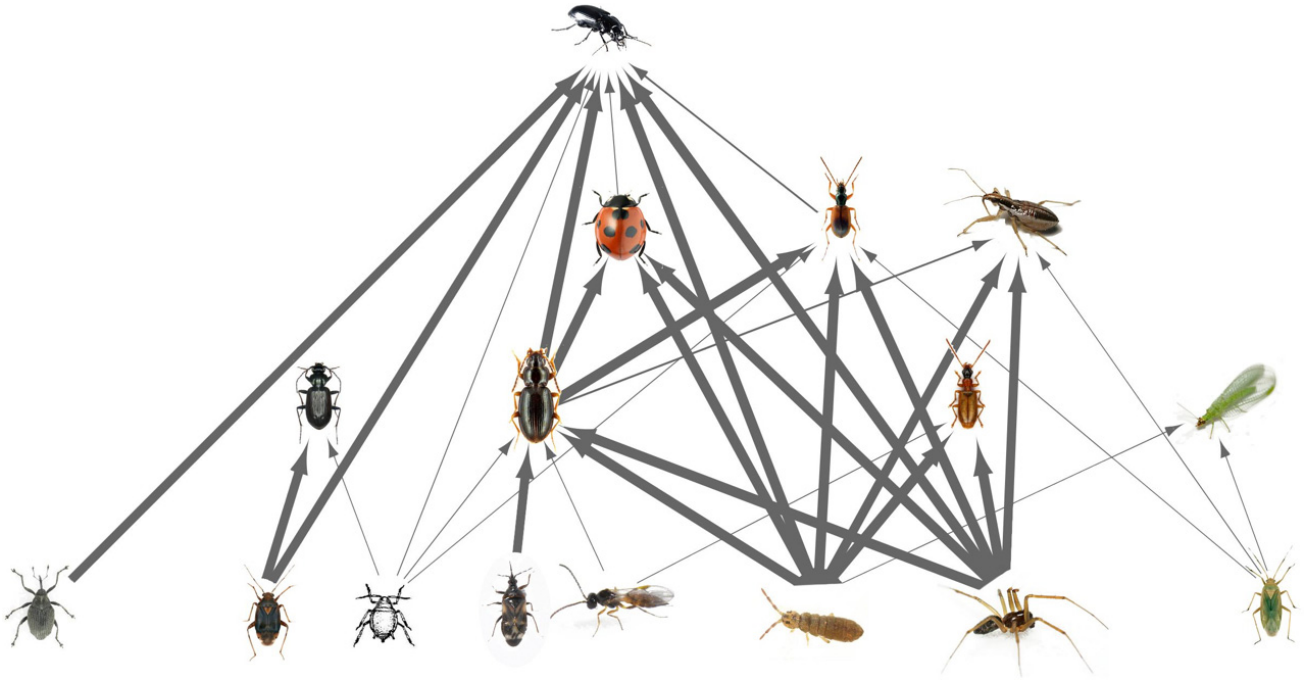

Figure 2. Trophic Representation of a Composite Food Web Reconstructed from Vortis Suction Sample Data Taken in 256 Agricultural Fields in the UK. The food web was learnt using the logical machine learning methods, as described in the text for the agro-ecological example and in Box 1. In place of learning species identity, here a functionally-based description of the data was adopted. The thickness of the trophic links represents the estimated frequency that a link was found in the data. The pictures for each functional node represents an archetypal species for that ecological function. Figure reproduced with permission from [28].

NGS describes a number of similar molecular technologies for generating large numbers of nucleic acid sequences for the identification of species (OTUs) and functions (Box 2). The great beauty of these methods is that the nucleic acids with which they work are common to all life forms and ubiquitous. In principle, NGS can be applied to the identification of OTUs and functions in environmental samples from any biome, habitat, and environment and any source material with minimal change in protocol. These benefits have driven the huge interest in eDNA as a source of data [31–33]. Our argument is that if coupled, machine learning and NGS data could serve as the foundation for a global, generic, and rapid network-based biomonitoring system that requires relatively little refinement to fit the environmental context in which it is deployed.

While the use of logic-based approaches to reconstructing networks from NGS data has still to be attempted, Weiss *et al.* [34] recently showed that statistical reconstruction of microbial networks from NGS data varied widely in sensitivity and precision. The reconstructed microbial networks often did not correspond well to those already understood. There is, therefore, room for improvement in statistical techniques of network reconstruction from NGS data [34]. However, in learning new structure in plankton networks, Lima-Mendez *et al.* [35] provide great encouragement that learning approaches can recover network information from NGS datasets.

Case Study – Reconstructing Networks from NGS Data

Jakuschkin *et al.* [36] have recently attempted this for microbes on the leaves of oak trees (*Quercus robur*) to identify potential antagonists of the causal agent of powdery-mildew, *Erysiphe alphitoides*. From DNA sampled from leaves with differing levels of symptoms, NGS identified OTUs that were used to generate a simple co-occurrence network [7] (Figure 3A), which revealed *E. alphitoides* to have numerous positive and negative associations that alter

Box 2. Next-Generation Sequencing (NGS)

Next-Generation Sequencing, or NGS, has become a prominent 'buzz-word' in ecological research, describing a number of distinct molecular-based methodologies. The two most common NGS methods are **amplicon** sequencing (metagenetics) and shotgun sequencing (metagenomics). These are used to examine the nucleic acids that are common to all life forms and by using this information identify taxa and their functions in samples from any biome, habitat and environment (e.g., water, soil, aerosols etc.), with minimal changes in methodology. Thus, NGS data can serve as the foundation for a generic, environment-independent, biomonitoring system able to resolve ecologically robust interaction networks by characterising the nodes (taxonomic species or OTUs) and links (functional interactions) present.

In NGS, DNA (or RNA) is first isolated from the sampled environment. Following DNA extraction [57], metagenetic approaches amplify phylogenetic or functional marker genes via **PCR**, targeted via a set of gene-specific primers, making sure only the genomic regions of interest are examined. 'Universal primers' exist for phylogenetically-informative marker genes from all three domains of life (e.g., 16S or 18S rRNA genes [57,58]). This approach can also be designed to target functional marker genes, notably those involved in biogeochemical cycling [59–63], providing assessments of ecosystem functioning. These amplicons of phylogenetic or functional marker genes are then sequenced [64]. Metagenomics omits the targeted amplification and instead, the extracted DNA is randomly fractured into small fragments covering all genes present [64]. These are then sequenced like the metagenetic amplicons. This avoids potential taxonomic biases that can occur in the metagenomics amplification step [65]. However, full gene coverage within a community requires considerable amounts of extracted DNA, and the costs of metagenomics is currently much higher than metagenetics.

NGS data uses differences across sequences to provide an estimate of taxonomic or functional diversity [66–68], and the number of identical sequences may also estimate relative abundance [69]. This is susceptible to primer biases, where the primers used preferentially amplify some OTUs at the expense of others. This bias can either be reduced via appropriate molecular approaches [70,71] or quantified to identify primers with no evidence of bias [65]. Furthermore, issues with sequencing errors, noise and statistical artefacts of the data (e.g., zero-rich data, non-normal distributions) have all been studied at length and appropriate bioinformatics approach exist to deal with these [66]. A major challenge is identifying the taxa and/or functions from the different DNA sequences present, by comparison with NGS sequence databases. While 16S rRNA gene databases are well populated with robust information, other gene databases can contain incorrectly assigned sequences or are missing data on entire taxonomic groups [43]. Thus, a significant global effort is required to provide robust, well-curated and maintained sequence databases.

the composition of the invaded community. The key question was whether these association links are functional and predictive of invasion and disease formation. Given there is no established background information in microbial network reconstruction for determining functional links, it was proposed that most of the co-occurrence links should be explained by environmental requirements shared between OTUs, with functional interactions persisting once the environmental correlations were accounted for [36]. The machine learning used a statistical inference approach, and this reconstructed a much smaller network of 26 OTU nodes, with fewer links (Figure 3B). Some of these links between the resident OTUs and *E. alphitoides* were suggestive of mechanisms that facilitate invasion and disease, while other interactions appeared suppressive.

Scientific Challenges

The NGS case study and the work of Lima-Mendez *et al.* [35] demonstrate clearly that NGS and learning approaches can generate hypotheses for ecological interactions, some of which have been validated [35]. Most importantly, however, this network proof of concept shows how rich biomonitoring would give insight into the ecosystem-wide biotic response to stressors. For learning methodologies to work well, they need data in the appropriate form and require guidance from background information that may not be readily available.

For the agro-ecological learning example, the data were transformed to a logarithmic ratio of abundance, while for the microbial network the OTU data came from across a gradient of infection. These learning datasets highlighted explicit contrasts, across either treatments or along gradients, in which the data structure exposed the linkages between the nodes of the network. In biomonitoring, no formal treatment structure will be present – the sampling is pure

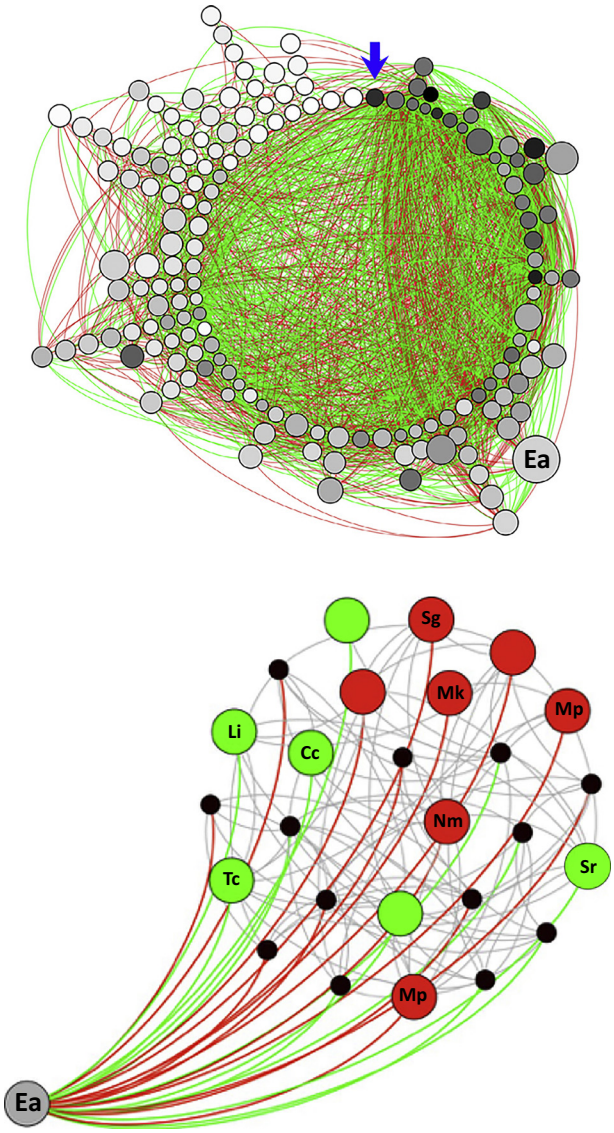

Figure 3. (A) Microbial Association Network of the Leaves of an Oak Tree (*Quercus robur* L.) Susceptible to the Foliar Fungal Pathogen, *Erysiphe alphitoides* (Ea). Each node represents a microbial taxon (either bacterial or fungal) and each link represents a significant correlation between their abundances. Red and green links indicate coexclusions and coassociations, respectively. The arrow indicates the node with the highest degree (i.e., the highest total number of links). Degree decreases clockwise, with nodes stacked on the same line having the same degree. The size of the nodes is inversely proportional to the sum of the correlation coefficients: larger nodes have more numbers and/or stronger negative associations. Darker nodes have higher betweenness centrality (calculated on the absolute values of associations), suggesting that they are topological keystone taxa. *E. alphitoides* is predominantly connected to the network through strong negative links (coexclusions) but is not a good candidate for topological keystone species. Figure reproduced with permission from [7]. (B) Network model of the pathobiome of *E. alphitoides* on oak leaves. Network nodes correspond to microbial OTUs. OTUs are linked if they are likely to interact together through direct ecological interactions, learnt using a Bayesian model of network inference. Small black nodes correspond to bacterial OTUs. The large nodes are fungal OTUs. Putative interactions between *E. alphitoides* and other OTUs are represented in green when the OTUs tend to facilitate each other and in red when they tend to exclude each other. Putative interactions between the interacting partners of *E. alphitoides* are shown in grey. The names of the fungal OTUs that could be assigned to species level are indicated. Abbreviations: Cc, *Cladosporium cladosporioides;* Li, *Lalaria inositophila;* Mk, *Monochaetia kansensis;* Mp, *Mycosphaerella punctiformis;* Nm, *Naevala minutissima;* Sg, *Sporobolomyces gracilis;* Sr, *Sporobolomyces roseus*; Tc, *Taphrina carpini.* Figure reproduced with permission from [36].

observation. Between replicate locations and over time, however, there will be natural perturbations of network structure. Appropriate binning/comparison across such variation in replicate NGS samples would therefore provide data of the correct structure for learning.

Declaring background information is more challenging, especially for systems where no clear rules have been established. Indeed, as a generic model for network structure, background information is something ecologists keenly want to know for all ecosystems. Models of allometric diet breadth or body size [14,37] might be tested for their genericity as background information on data from relatively unstudied systems. The microbial case study suggests that although using environmental requirements as background information is a weak model, it could nonetheless generate adequate networks for detecting environmental change and biomonitoring. Moreover, new developments in machine learning, such as meta-interpretative learning (MIL) [38,39], suggest that powerful background information can be discovered from data. This was demonstrated by discovering the rule that 'big things eat small things' directly from data for a simulated, synthetic food web [40]. Reconstructing of an ever greater number of ecological networks will drive an improvement in our understanding of ecosystem structure and function, and further generate better background information and ecological knowledge.

Next-Generation Global Biomonitoring (NGB)

The amalgam of machine learning and NGS for biomonitoring requires a technological framework in which to work. In Figure 1, we outline our concept for next-generation global biomonitoring, which unites this science with advances in technology, particularly of miniaturisation, communications, and cloud-based storage and processing. We envision the development of an automated sampler comprised of key technological components that already exist, but which have yet to be combined. Miniaturisation of NGS sequencing machines has already produced a sequencer that is the size of a USB stick and can run samples anywhere for many days at a time[i]. DNA extraction, which is currently one of the most time-consuming steps in environmental molecular ecology, has the potential to be fully automated via robotic liquid handling platforms that could ultimately be miniaturised and deployed remotely (e.g., [41]). Sample preparation equipment is also being developed.

Combining samplers and sequencers will form the basis of an autonomous sampler, with a known duration of operation. Many thousands of units could be deployed in different ecosystems across a global array at a fraction of the costs of traditional biomonitoring. Being remote, it is necessary to build-in appropriate intelligence to allow the device to act autonomously and to communicate data. This can be done using off-the-shelf mobile communications modules that automatically location- and time-stamp the data using GPS. Mesh-network solutions, such as LoRa[ii] can create communication networks at very large grain (~10 km between samplers) and only when it is necessary to transfer data, providing considerable power and cost savings over current 3G/4G mobile telephone and satellite networks. An embedded processor would then manage the process of sampling, sequencing, and uploading of sequence data, and a battery and solar panels would supply electrical power. Less complex hand-held samplers might fill gaps in the global array or provide information for specific end-users. For example, farmers could sample their fields for crop pathogens or medical technicians could sample for the presence and evolution of all diseases, much as is currently being done using NGS for Ebola alone [42]. Drones might also be used to more cheaply collect and return samples to the laboratory for processing. Importantly, whether from a hand-held, drone or remote sampler, the sequence data should be of the same quality, sharing common date and location stamping. The uploaded data would be stored in a cloud database, where it would be collated and checked prior to being linked with online sequence databases to automate detection of the OTUs or functions present. Finally, machine-learning would reconstruct the networks implicit in the sample data and determine potential changes in structure in real time. It is only at this point,

once change in an ecosystem has been detected, that humans would explicitly enter the NGB process to interpret the ecology.

This infrastructure would not need to be created anew for NGB: very large databases are already curated in the cloud and machine-learning approaches serve information to end-users via the internet. The proof that much of this technology exists right now is in our pockets. Smartphones combine embedded processing, GPS, and communications, and communicate with online machine learning structures, such as Google Translate[iii] and Apple Siri[iv]. The large-scale storage and processing needed for NGB could 'piggy-back' on the existing infrastructure used in remote sensing. Indeed, one of the potentially profound results of NGB would be to couple ecological biomonitoring to large-scale remote sensing, giving better understanding of environmental and ecological change.

Technological Challenges

There are three major technological barriers to this NGB approach. The first is the sample mechanism of the automated sampler that should deliver an uncontaminated sample from the environment being biomonitored. Such samplers already exist for liquid water and air environments. For other environments, such as in the soil or at the soil surface, work will be necessary to construct a sample mechanism. As with all ecological protocols, though, the sample mechanism will set limits on those OTUs that can be sampled and reassembled into a network. OTUs not sampled automatically might be inferred from data, such as remote sensing.

The second technological barrier is the quality of current OTU databases. While certain gene databases are reasonably well populated and can return robust information, the coverage of all genes and taxa is still incomplete and databases contain incorrectly assigned sequences or omit entire taxonomic groups [43]. A significant, global effort is required to provide robust, well-curated and maintained sequence databases for use in NGB; fortunately, this is already underway for many taxonomic groups (e.g., International Barcode of Life[v]). NGB could accelerate this process by providing the 'big picture' impetus for generating shared sequence databases that would transform NGS, but which, because of time and cost constraints, are still lacking.

The third technological barrier is the lack of a statistical framework for sampling using networks. While power analysis to detect a given ecological effect is well established in biomonitoring [44], for NGB approaches the statistics of the size of the array of samplers necessary to detect network change in real time will need to be addressed. Similarly, over a period of calibration, the natural variation in network structure will need to be estimated as a baseline against which change can be detected/tested.

Conclusion: Potential Benefits of NGB

The costs of building an NGB array will be considerable, but much of the infrastructure and technology already exists and the price of the sampler and sampling will likely fall with economies of scale. As the system is largely autonomous, humans only enter the biomonitoring process once the data are acquired, thus removing the high fixed costs of labour-intensive biomonitoring.

Real-time NGB would be far more sensitive than current approaches, and the window for identifying when a stressor first elicits a response could be greatly foreshortened. The approach solves the three key problems, of accuracy, costs and generality of current approaches to biomonitoring. By broadening biomonitoring from single indicators, the approach enrichens the evaluation of change in ecosystem structure and function. The costs of time and effort in reconstructing networks and, as a consequence, detecting change in an ecosystem are

Outstanding Questions

Scientific challenges: The scientific challenges in NGB lie in: (i) understanding and comparing the different methods of learning, so that the appropriate method may be selected for a particular learning setting; (ii) developing an understanding of how to appropriately treat NGS datasets so that the interaction linkages to be discovered are best 'exposed' for learning; (iii) testing of current models for network structure as background knowledge for generic reconstruction of ecological networks from known and unknown ecosystems; and (iv) research to refine our current understanding of how network and ecosystem structure lead to function, for interpretation of ecosystem change within the current biomonitoring framework.

Technical challenges: The current limitations on NGB largely lie in (i) building sampling mechanisms that can work in different environments (e.g., suction sample apparatus for sampling the biodiversity of the air and modified wet pitfall traps for sampling ground dwelling organisms); (ii) improving the quality of OTU databases, through a concerted global programme; and (iii) developing a statistical framework to detect and biomonitor ecosystem change in large-scale, replicated networks.

Policy Challenges: Ecological networks, per se, have no position in national or international policy and decision-making. The shift to thinking about the management of natural resources as ecosystem services that can interact creates an opportunity to insert network approaches into policy. Recently, there have been attempts to address this problem by fusing ecological networks to social and economic networks [22,23]. It is possible that NGB could provide the ecological network component in such approaches.

Funding: To date the different elements of this work have proceeded in parallel and in a rather uncoordinated manner, funded through local, national, and industrial funding mechanisms. The development of NGB will require more concerted international funding, guided by already existing bodies, including the Group on Earth Observations (GEO) Initiatives (e.g., GEOSS, GEOGLAM, GEOBON, GFOI,

markedly reduced. In turn, NGB is general because it is based on sampling ubiquitous nucleic acids.

Importantly, fusing NGS and machine learning allows us to learn ecology. The agro-ecological example demonstrates that machine learning can be used for 'hypothesis-test' science. With this fusion, it should be possible to reconstruct networks for ecosystems about which we have very little understanding and discover novel interactions within the ecosystems we already work in. Moreover, this large-scale biomonitoring will likely work well with existing large-scale monitoring approaches, such as remote sensing of the Earth's environments.

Traditionally, there has been relatively little exchange and cross-fertilization between the disciplines of biomonitoring and ecology. A shift towards a large-scale biomonitoring approach that measures change in terms of network structure would provide richer, more ecological information that is moreover comparable between ecosystems. NGB data would add to our knowledge explicitly and foster a revolution in ecological understanding of ecosystem change.

Acknowledgements

Dave Bohan and Corinne Vacher receive support from INRA MEM project Learn-Biocontrol. Support is also provided to Dave Bohan from the ANR FACCE JPI SURPLUS PREAR project. Alireza Tamaddoni-Nezhad, Alex Dumbrell and Guy Woodward are funded by the NE/M020843/1 and NE/M02086X/1. We acknowledge past support from the Imperial College – Syngenta Ltd UIC.

Resources

[i]MinION https://www.nanoporetech.com

[ii]https://www.lora-alliance.org

[iii]https://backchannel.com/how-google-is-remaking-itself-as-a-machine-learning-first-company-ada63defcb70#.fnm9q9pnq

[iv]http://www.macrumors.com/2016/08/24/apple-machine-learning-siri-neural-network/

[v]http://ibol.org

References

1. Millennium Ecosystem Assessment (2005) *Ecosystems and Human Well-Being: Synthesis,* Island Press
2. Díaz, S. *et al.* (2015) The IPBES Conceptual Framework – connecting nature and people. *Curr. Opin. Environ. Sustain.* 14, 1–16
3. Smodiš, B. *et al.* (1999) Biomonitoring of air pollution as exemplified by recent IAEA programs. *Biol. Trace Elem. Res.* 71, 257–266
4. Karr, J.R. (1981) Assessment of biotic integrity using fish communities. *Fisheries* 6, 21–27
5. Girardin, P. *et al.* (1999) Indicators: tools to evaluate the environmental impacts of farming systems. *J. Sustainable Agric.* 13, 5–21
6. Niemeijer, D. (2002) Developing indicators for environmental policy: data-driven and theory-driven approaches examined by example. *Env. Sci. Policy* 5, 91–103
7. Vacher, C. *et al.* (2016) Learning ecological networks from next-generation sequencing data. *Adv. Ecol. Res.* 54, 1–39
8. Blanchard, J.L. (2015) Climate change: a rewired food web. *Nature* 527, 173–174
9. Thompson, M.S.A. *et al.* (2015) Gene-to-ecosystem impacts of a catastrophic pesticide spill: testing a multilevel bioassessment approach in a river ecosystem. *Freshwater Biol.* 61, 2037–2050
10. Aizen, M.A. *et al.* (2008) Invasive mutualists erode native pollination webs. *PLoS Biol.* 6, e31
11. Thompson, R.M. *et al.* (2012) Food webs: reconciling the structure and function of biodiversity. *Trends Ecol. Evol.* 27, 689–697
12. Pocock, M.J.O. *et al.* (2012) The robustness and restoration of a network of ecological networks. *Science* 335, 973–977
13. Montoya, J.M. *et al.* (2006) Ecological networks and their fragility. *Nature* 442, 259–264
14. Petchey, O.L. *et al.* (2008) Size, foraging, and food web structure. *Proc. Natl. Acad. Sci. U. S. A.* 105, 4191–4196
15. Pan, Q. *et al.* (2016) Effects of functional diversity loss on ecosystem functions are influenced by compensation. *Ecology* 97, 2293–2302
16. Frost, C.M. *et al.* (2016) Apparent competition drives community-wide parasitism rates and changes in host abundance across ecosystem boundaries. *Nat. Commun.* 7, 12644
17. Layer, K. *et al.* (2011) Long-term dynamics of a well-characterised food web. *Adv. Ecol. Res.* 44, 69–117
18. Carpenter, S.R. *et al.* (2001) Trophic cascades, nutrients, and lake productivity: whole-lake experiments. *Ecol. Monogr.* 71, 163–186
19. Scheffer, M. *et al.* (2001) Catastrophic shifts in ecosystems. *Nature* 413, 591–596
20. Beisner, B.E. *et al.* (2003) Alternative stable states in ecology. *Front. Ecol. Environ.* 1, 376–382
21. Raffaelli, D. (2006) Food webs, body size and the curse of the Latin binomial. In *Energetics to Ecosystems: The Dynamics and Structure of Ecological Systems* (Rooney, N. *et al.*, eds), Springer, pp. 53–64
22. QUINTESSENCE Consortium (2016) Networking our way to better ecosystem service provision. *Trends Ecol. Evol.* 31, 105–115
23. Dee, L.E. *et al.* (2016) Operationalizing network theory for ecosystem service assessments. *Trends Ecol. Evol.* 32, 118–130
24. Haughton, A.J. *et al.* (2003) Invertebrate responses to the management of genetically modified herbicide-tolerant and conventional spring crops. II. Within-field epigeal and aerial arthropods. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 358, 1863–1877

GMOS, AFRIGEOSS, BLUE PLANET) and the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES).

25. Bohan, D.A. *et al.* (2005) Effects on weed and invertebrate abundance and diversity of herbicide management in genetically modified herbicide-tolerant winter-sown oilseed rape. *Proc. R. Soc. Biol. Sci. Ser. B* 272, 463–474
26. Bohan, D.A. *et al.* (2011) Automated discovery of food webs from ecological data using logic-based machine learning. *PLoS One* 6, e29028
27. Muggleton, S.H. *et al.* (2000) Theory completion using inverse entailment. *Proceedings of the 10th International Workshop on Inductive Logic Programming* Springer, pp. 130–146
28. Tamaddoni-Nezhad, A. *et al.* (2013) Construction and validation of food webs using logic-based machine learning and text mining. *Adv. Ecol. Res.* 49, 225–289
29. Tamaddoni-Nezhad, A. *et al.* (2012) Machine learning a probabilistic network of ecological interactions. *Proceedings of the 21st International Conference on Inductive Logic Programming, LNAI 7207* Springer, pp. 332–346
30. Davey, J.S. *et al.* (2013) Intraguild predation in winter wheat: prey choice by a common epigeal carabid consuming spiders. *J. Appl. Ecol.* 50, 271–279
31. Barnes, M.A. and Turner, C.R. (2015) The ecology of environmental DNA and implications for conservation genetics. *Conserv. Genet.* 17, 1–17
32. Thomsen, P.F. and Willerslev, E. (2015) Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18
33. Evans, D.M. *et al.* (2016) Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Funct. Ecol.* 30, 1904–1916
34. Weiss, S. *et al.* (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681
35. Lima-Mendez, G. *et al.* (2015) Determinants of community structure in the global plankton interactome. *Science* 348, 1262073–1262073
36. Jakuschkin, B. *et al.* (2016) Deciphering the pathobiome: intra- and inter-kingdom interactions involving the pathogen *Erysiphe alphitoides*. *Microb. Ecol.* 72, 870–880
37. Woodward, G. *et al.* (2010) Individual-based food webs: species identity, body size and sampling effects. *Adv. Ecol. Res.* 43, 211–266
38. Muggleton, S.H. *et al.* (2014) Meta-interpretive learning: application to grammatical inference. *Mach. Learn.* 94, 25–49
39. Muggleton, S.H. *et al.* (2015) Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited. *Mach. Learn.* 100, 49–73
40. Tamaddoni-Nezhad, A. *et al.* (2015) Towards machine learning of predictive models from ecological data. *Proceedings of the 24th International Conference on Inductive Logic Programming* Springer-Verlag, pp. 154–167
41. Anderson, R.C. *et al.* (2000) A miniature integrated device for automated multistep genetic assays. *Nucleic Acids Res.* 28, e60
42. Quick, J. *et al.* (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232
43. Tedersoo, L. *et al.* (2011) Tidying up international nucleotide sequence databases: ecological, geographical and sequence quality annotation of ITS sequences of mycorrhizal fungi. *PLoS One* 6, e24940
44. Osenberg, C.W. *et al.* (1994) Detection of environmental impacts: natural variability, effect size, and power analysis. *Ecol. Appl.* 4, 16–30
45. Stein, R.R. *et al.* (2013) Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.* 9, 31–36
46. Faust, K. *et al.* (2015) Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66
47. Bucci, V. *et al.* (2016) MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol.* 17, 121
48. Friedman, J. and Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8, e1002687
49. Li, C. *et al.* (2016) Predicting microbial interactions through computational approaches. *Methods* 102, 12–19
50. Kurtz, Z.D. *et al.* (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11, e1004226
51. Faisal, A. *et al.* (2010) Inferring species interaction networks from species abundance data: a comparative evaluation of various statistical and machine learning methods. *Ecol. Inform.* 5, 451–464
52. Cazelles, K. *et al.* (2016) A theory for species co-occurrence in interaction networks. *Theor. Ecol.* 9, 39–48
53. Ovaskainen, O. *et al.* (2016) Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* 7, 549–555
54. Muggleton, S. (1991) Inductive logic programming. *New Generat. Comput.* 8, 295–318
55. Tamaddoni-Nezhad, A. *et al.* (2006) Application of abductive ILP to learning metabolic network inhibition from temporal data. *Mach. Learn.* 64, 209–230
56. King, R.D. *et al.* (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427, 247–252
57. Carugati, L. *et al.* (2015) Metagenetic tools for the census of marine meiofaunal biodiversity: an overview. *Mar. Genomics* 24, 11–20
58. Lanzén, A. *et al.* (2016) Multi-targeted metagenetic analysis of the influence of climate and environmental parameters on soil microbial communities along an elevational gradient. *Sci. Rep.* 6, 28257
59. Barbi, F. *et al.* (2014) PCR primers to study the diversity of expressed fungal genes encoding lignocellulolytic enzymes in soils using high-throughput sequencing. *PLoS One* 9, e116264
60. Gaby, J.C. *et al.* (2012) A comprehensive evaluation of PCR primers to amplify the nifH gene of nitrogenase. *PLoS One* 7, e42149
61. Lansdown, K. *et al.* (2016) Importance and controls of anaerobic ammonium oxidation influenced by riverbed geology. *Nat. Geosci.* 9, 357–360
62. Li, J. *et al.* (2015) amoA gene abundances and nitrification potential rates suggest that benthic ammonia-oxidizing bacteria (AOB) not archaea (AOA) dominate N cycling in the Colne estuary, UK. *Appl. Environ. Microbiol.* 81, 159–165
63. Papaspyrou, S. *et al.* (2014) Nitrate reduction functional genes and nitrate reduction potentials persist in deeper estuarine sediments. Why? *PLoS One* 9, e94111
64. D'Amore, R. *et al.* (2016) A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* 17, 2937
65. Cotton, T.E.A. *et al.* (2014) What goes in must come out: testing for biases in molecular analysis of arbuscular mycorrhizal fungal communities. *PLoS One* 9, e109234
66. Dumbrell, A.J. *et al.* (2017) Microbial community analysis by single-amplicon high-throughput next generation sequencing: data analysis – from raw output to ecology. In *Hydrocarbon and Lipid Microbiology Protocols* (McGenity, T.J. *et al.*, eds), Springer, pp. 155–206
67. Caporaso, J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336
68. Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998
69. Thomas, T. *et al.* (2012) Metagenomics – a guide from sampling to data analysis. *Microb. Inform. Exp.* 2, 3
70. Polz, M.F. and Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* 64, 3724–3730
71. Sipos, R. *et al.* (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* 60, 341–350